


BMJ Open Reducing the number of unnecessary biopsies of US-BI-RADS 4a lesions through a deep learning method for residents-in-training: a cross-sectional study

Chenyang Zhao , Mengsu Xiao, He Liu, Ming Wang, Hongyan Wang, Jing Zhang, Yuxin Jiang, Qingli Zhu

To cite: Zhao C, Xiao M, Liu H, *et al.* Reducing the number of unnecessary biopsies of US-BI-RADS 4a lesions through a deep learning method for residents-in-training: a cross-sectional study. *BMJ Open* 2020;**10**:e035757. doi:10.1136/bmjopen-2019-035757

► Prepublication history for this paper is available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2019-035757>).

Received 14 November 2019
Revised 01 April 2020
Accepted 28 April 2020



© Author(s) (or their employer(s)) 2020. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

Department of Ultrasound, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

Correspondence to

Dr Qingli Zhu;
zhuqingli@pumch.cn and
Professor Yuxin Jiang;
jiangyuxinxh@163.com

ABSTRACT

Objective The aim of the study is to explore the potential value of S-Detect for residents-in-training, a computer-assisted diagnosis system based on deep learning (DL) algorithm.

Methods The study was designed as a cross-sectional study. Routine breast ultrasound examinations were conducted by an experienced radiologist. The ultrasonic images of the lesions were retrospectively assessed by five residents-in-training according to the Breast Imaging Report and Data System (BI-RADS) lexicon, and a dichotomic classification of the lesions was provided by S-Detect. The diagnostic performances of S-Detect and the five residents were measured and compared using the pathological results as the gold standard. The category 4a lesions assessed by the residents were downgraded to possibly benign as classified by S-Detect. The diagnostic performance of the integrated results was compared with the original results of the residents.

Participants A total of 195 focal breast lesions were consecutively enrolled, including 82 malignant lesions and 113 benign lesions.

Results S-Detect presented higher specificity (77.88%) and area under the curve (AUC) (0.82) than the residents (specificity: 19.47%–48.67%, AUC: 0.62–0.74). A total of 24, 31, 38, 32 and 42 identified as BI-RADS 4a lesions by residents 1, 2, 3, 4 and 5 were downgraded to possibly benign lesions by S-Detect, respectively. Among these downgraded lesions, 24, 28, 35, 30 and 40 lesions were proven to be pathologically benign, respectively. After combining the residents' results with the results of the software in category 4a lesions, the specificity and AUC of the five residents significantly improved (specificity: 46.02%–76.11%, AUC: 0.71–0.85, $p < 0.001$). The intraclass correlation coefficient of the five residents also increased after integration (from 0.480 to 0.643).

Conclusions With the help of the DL software, the specificity, overall diagnostic performance and interobserver agreement of the residents greatly improved. The software can be used as adjunctive tool for residents-in-training, downgrading 4a lesions to possibly benign and reducing unnecessary biopsies.

Strengths and limitations of this study

- We focused on the value of the deep learning software in helping inexperienced ultrasound (US) readers, with a total of five residents involved in the study.
- The Breast Imaging Report and Data System (BI-RADS) 4a lesions, where unnecessary biopsies are often performed and which pose clinical difficulty among US readers, were selected to further explore the potential of the computer-aided diagnosis (CAD) system.
- The US operator and five readers all received training on BI-RADS lexicon and usage of the CAD system for this study, making the results more reliable.
- Because only static images were provided, the performance of the residents could be underestimated.
- The study was conducted in a single centre and might not reflect real clinical settings in other medical centres.

INTRODUCTION

On account of the increasing incidence rate in the past decade, breast cancer has become a growing public health concern worldwide.^{1 2} Early detection of breast cancer can largely improve patient prognosis.^{3–5} As an important adjunctive tool to mammography, ultrasound (US) has shown great potential for diagnosing breast masses, especially in dense breast tissue, allowing identification of masses that are occult on mammography.⁶ Considering its accessibility and cost-effectiveness, US has become the most popular imaging method for breast screening in China and has also been proven to be superior or not inferior to mammography in diagnostic performance.⁷

Nevertheless, low specificity and high interobserver variability remain problematic

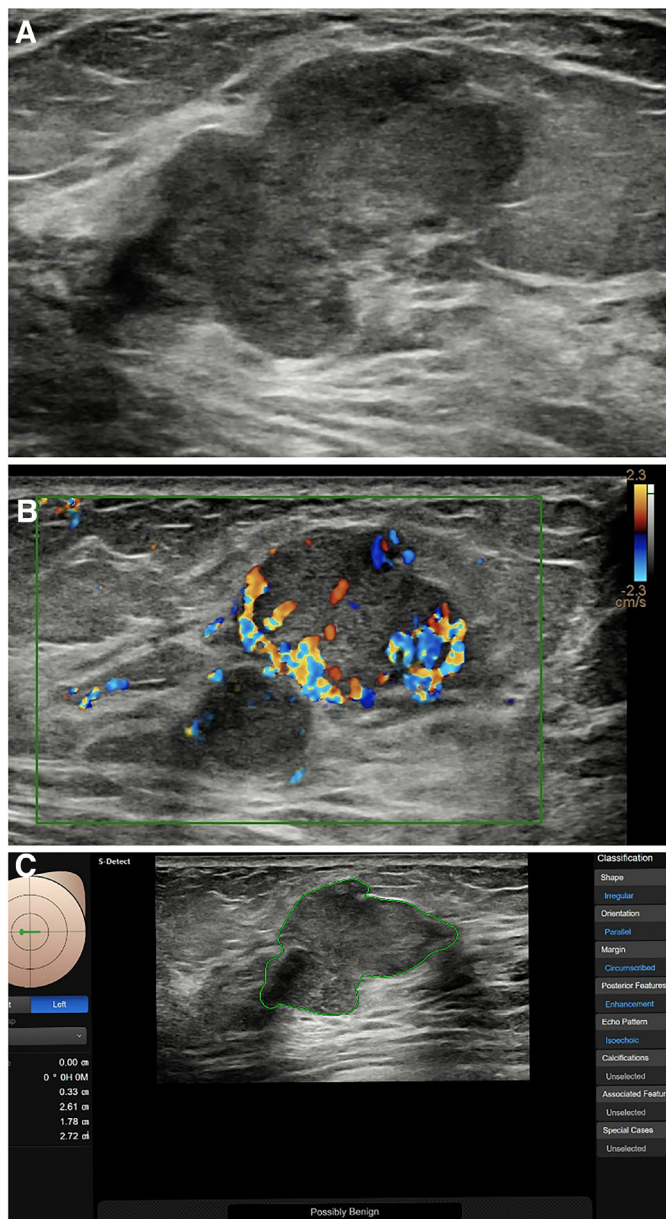


Figure 1 Example of a downgraded 4a lesion. The hypoechoic lesion with slightly irregular shape and abundant vascularity was classified into a 4a lesion by four residents and the software diagnosed it as a possibly benign one. The mass was verified as a benign phyllodes tumour on histopathology. (A) The section in the maximal size of the lesion. (B) Colour Doppler imaging of the lesion (the section vertical to A). (C) The working interface of S-Detect.

disadvantages of US, especially for residents who received only short-term training in breast US.^{8–11} Although the Breast Imaging Report and Data System (BI-RADS) lexicon was proposed by the American College of Radiology,^{12–14} residents-in-training are still inclined to have relatively poor diagnostic performance when assessing breast lesions.¹⁵ According to the lexicon, the features of breast lesions that suggest malignancy include irregular shape, unparallel orientation, indistinct/angular/microlobulated/spiculated margin, echogenic halo and microcalcifications. Solid masses with slightly abnormal shape or

margins but no other malignant evidence are categorised as BI-RADS 4a. These 4a lesions, which present a few low-level suspicious features but mainly benign characteristics, can create confusion among inexperienced residents during lesion classification and can result in wrong judgements, subsequently leading to overtreatment. It has been illustrated by previous studies that the rate of malignancy of BI-RADS 4a lesions was 3%–10%, most of which were benign lesions but received unnecessary biopsies or surgery. It is worth developing new techniques with higher specificity than conventional methods to address this issue.¹⁶

Computer-aided diagnosis (CAD) systems have played a growing part in many fields of medical imaging, including breast US.^{17–21} S-Detect for Breast is a cutting-edge CAD system that acts as adjunctive tool for US imaging diagnosis of breast lesions. Unlike conventional CAD systems for medical imaging, it was developed based on a deep learning algorithm, which was constructed on the convolutional neural network and learnt from large quantities of ultrasonic images. The diagnostic efficacy of the CAD software in classifying breast lesions has been validated by several studies.^{22–24} Furthermore, S-Detect has been proven to be of value in increasing the diagnostic performance of residents-in-training.^{22–25} The possibly benign BI-RADS 4a lesions posed a potential challenge for breast US, which could be especially difficult for inexperienced residents-in-training. Also, as far as we know, the feasibility of S-Detect in improving the diagnostic accuracy of residents-in-training in detecting BI-RADS 4a lesions has not been investigated in previous studies.

In this study, we evaluated the diagnostic performance of S-Detect and five residents-in-training in classifying breast lesions. The results of the residents were re-evaluated after some of the category 4a lesions were downgraded by CAD. The aim of the study was to further explore the potential role of S-Detect in aiding in-training readers and determine how this system can help improve diagnostic performance, especially for BI-RADS category 4a lesions.

MATERIALS AND METHODS

This study was a cross-sectional observational study. Written informed consent was obtained from the adult patients of the study. For patients under 18 years old, written informed consent was signed by their guardians who accompanied them during the US examination.

Patients

Patient and public involvement

There was no direct patient involvement in this study. No patients were involved in the study design and in the writing or editing of the article. During the examination process, we introduced the study outline and the new imaging software to each patient who participated. Patients or the public were not involved in the design,

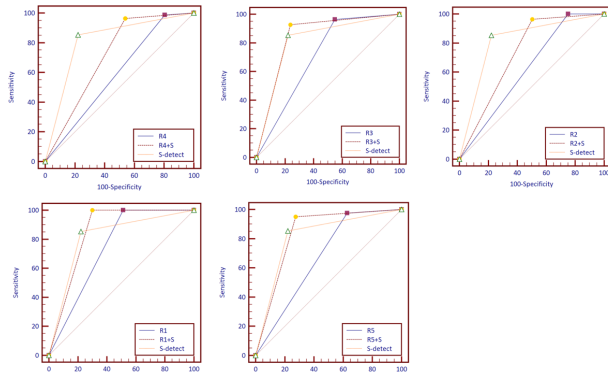


Figure 2 The receiver operating characteristics curve of the five residents (R), S-Detect and the integrated results of residents and S-Detect.

or conduct, or reporting, or dissemination plans of our research.

Patient recruitment

A total of 195 focal breast lesions from patients aged between 15 and 82 years from July 2018 to March 2019, with a mean age of 45.7 years and a median of 45.0 years, were enrolled consecutively in this study.

The inclusion criteria for the study were as follows:

- ▶ Palpable masses verified by breast imaging.
- ▶ Non-palpable masses found by breast imaging, with or without other symptoms.

The exclusion criteria were as follows:

- ▶ Biopsy of the breast lesions performed before US examinations.
- ▶ Pregnancy or in lactation.
- ▶ Neoadjuvant treatment.

- ▶ Only simple cysts visible on US images.
- ▶ No evident focal breast lesions suitable for CAD evaluation.

The patients underwent US examinations before they received further treatment. All lesions underwent core biopsy or surgery or a combination of these. Any malignant, atypical or high-risk core biopsy result (including lobular carcinoma in situ, atypical ductal hyperplasia, papillary lesion) prompted excision. The final pathological diagnosis was made by a senior pathologist (with 20 years of experience in breast pathology). The pathological results were deemed as the gold standard in this study.

Study protocol

Image assessment of S-Detect for Breast and the five residents-in-training

The patients received standard bilateral breast US scans performed by an experienced radiologist (with 15 years of experience in breast US). A commercial US unit (RS85, Samsung Medison, Korea) equipped with an L3-12A high-frequency linear probe (3–12 MHz) and the CAD software S-Detect for Breast (Samsung Healthcare, South Korea) were used to perform greyscale US, colour Doppler US and strain elastography.

At least two typical images of longitudinal and sectional greyscale US, colour Doppler and elastography of each lesion were recorded for further evaluation.

A single greyscale US image demonstrating the lesion with the maximum size was manually selected for S-Detect breast analysis. First, the radiologist clicked the centre of the target mass, and the contour of the lesion was automatically segmented by S-Detect. The outline of the lesion

Table 1 Pathological results and BI-RADS classifications of the breast lesions

| Pathological results | | BI-RADS classifications | | | | |
|--------------------------------------|------------|-------------------------|-----|-----|-----|-----|
| | n (%) | R1 | R2 | R3 | R4 | R5 |
| Malignant lesions | | | | | | |
| Intraductal carcinoma | 7 (8.54) | | | | | |
| Invasive ductal carcinoma | 66 (80.49) | | | | | |
| Invasive lobular carcinoma | 4 (4.88) | | | | | |
| Neuroendocrine intraductal carcinoma | 2 (2.44) | | | | | |
| Invasive micropapillary carcinoma | 1 (1.22) | | | | | |
| Mucinous carcinoma | 2 (2.44) | | | | | |
| Total | 82 | | | | | |
| Benign lesions | | | | | | |
| Adenosis | 18 (15.93) | | | | | |
| Intraductal papillomas | 12 (10.62) | | | | | |
| Lobular tumour | 2 (1.77) | | | | | |
| Chronic inflammation | 4 (3.54) | | | | | |
| Adiponecrosis | 1 (0.88) | | | | | |
| Total | 113 | | | | | |
| | | 3 | 4a | 4b | 4c | 5 |
| | | 55 | 32 | 43 | 59 | 6 |
| | | 28 | 37 | 44 | 75 | 11 |
| | | 54 | 52 | 51 | 29 | 9 |
| | | 23 | 39 | 51 | 60 | 22 |
| | | 44 | 56 | 64 | 26 | 5 |
| Total | | 195 | 195 | 195 | 195 | 195 |

BI-RADS, Breast Imaging Report and Data System; R, resident.

was adjusted manually by the radiologist when necessary. Then, the classification of each lesion in a dichotomic form (possibly benign and possibly malignant) was provided by S-Detect. The extracted US descriptors were also displayed, including shape, orientation, margins, pattern and posterior acoustic features.

Five residents-in-training with 1–3 years of working experience were invited to assess the US lesions independently. All images of the lesions (including greyscale, colour Doppler flow and elastography images) were retrospectively reviewed by the five residents-in-training. According to BI-RADS lexicon, irregular shape, unparallel orientation, indistinct/angular/microlobulated/spiculated margin, echogenic halo and microcalcifications were considered as malignant greyscale US features. For strain elastography, elasticity scores of a 5-point scale based on colour mapping from red (soft), to green (intermediate), to blue (hard) were assessed.^{26 27} Abundant vascularity and hard elasticity (elasticity scores of 4–5) were considered malignant features. After identifying the ultrasonographic features, the five residents classified

those lesions into BI-RADS 3, 4a, 4b, 4c and 5. The residents were blinded to S-Detect and pathology results. R1–R5 were used to represent the five residents. R1, R2 and R3 were third-year residents, and each had 1-year experience with breast US. R4 and R5 were second-year residents, each with 6 months of experience with breast US. All five residents had received a standard training programme for breast US, and they have also passed the examinations for basic US organised by our medical centre.

A cut-off value was set at category 4 to transform the residents' results into a dichotomic form. Category 2 and 3 lesions were deemed as possibly benign, and category 4 and 5 were considered possibly malignant. The diagnostic performances of S-Detect and the five residents were evaluated, and comparisons were made between S-Detect and the residents.

Integration of the results of the five residents and S-Detect for Breast

To evaluate the potential of S-Detect in helping improve the diagnostic accuracy of residents, the results of the five

Table 2 Diagnostic performance of S-Detect, the five residents and the integrated results

| | SE (%) | SP (%) | PLR | NLR | PPV (%) | NPV (%) | AUC |
|----------|----------------|----------------|--------------|--------------|----------------|-----------------|--------------|
| | 95% CI | 95% CI | 95% CI | 95% CI | 95% CI | 95% CI | 95% CI |
| S-Detect | 85.37 | 77.88 | 3.86 | 0.19 | 73.68 | 88 | 0.82 |
| | 75.83 to 92.20 | 69.10 to 85.14 | 2.70 to 5.52 | 0.11 to 0.32 | 63.65 to 82.19 | 79.98 to 93.64 | 0.75 to 0.87 |
| R1 | 100 | 48.67 | 1.95 | 0 | 58.57 | 100 | 0.74 |
| | 95.60 to 100 | 39.16 to 58.26 | 1.63 to 2.33 | 0 | 49.95 to 66.83 | 87.66 to 100.00 | 0.68 to 0.80 |
| R2 | 100 | 24.78 | 1.33 | 0 | 49.1 | 100 | 0.62 |
| | 95.60 to 100 | 17.14 to 33.78 | 1.20 to 1.48 | 0 | 41.30 to 56.94 | 87.66 to 100.00 | 0.55 to 0.69 |
| R3 | 96.34 | 45.13 | 1.76 | 0.08 | 56.03 | 94.44 | 0.71 |
| | 89.68 to 99.24 | 35.75 to 54.77 | 1.48 to 2.09 | 0.03 to 0.25 | 47.43 to 64.37 | 84.61 to 98.84 | 0.64 to 0.77 |
| R4 | 98.78 | 19.47 | 1.23 | 0.06 | 47.09 | 95.65 | 0.59 |
| | 93.39 to 99.97 | 12.62 to 27.98 | 1.12 to 1.35 | 0.01 to 0.46 | 39.45 to 54.84 | 78.05 to 99.89 | 0.52 to 0.66 |
| R5 | 97.56 | 37.17 | 1.55 | 0.07 | 52.98 | 95.45 | 0.67 |
| | 92.47 to 99.70 | 28.26 to 46.76 | 1.34 to 1.80 | 0.02 to 0.26 | 44.70 to 61.14 | 84.53 to 99.44 | 0.60 to 0.74 |
| R1+S | 100 | 69.91* | 3.32 | 0 | 70.69 | 100 | 0.85* |
| | 95.60 to 100 | 60.57 to 78.18 | 2.51 to 4.40 | 0 | 61.52 to 78.77 | 95.44 to 100 | 0.79 to 0.90 |
| R2+S | 96.34* | 49.56* | 1.91 | 0.07 | 58.09 | 94.92 | 0.73* |
| | 89.68 to 99.24 | 40.02 to 59.12 | 1.58 to 2.30 | 0.02 to 0.23 | 49.33 to 66.49 | 85.85 to 98.94 | 0.66 to 0.79 |
| R3+S | 92.68* | 76.11* | 3.88 | 0.1 | 73.79 | 93.48 | 0.84* |
| | 84.75 to 97.27 | 67.17 to 83.63 | 2.78 to 5.42 | 0.04 to 0.21 | 64.20 to 81.96 | 86.34 to 97.57 | 0.79 to 0.89 |
| R4+S | 96.34† | 46.02* | 1.78 | 0.08 | 56.43 | 94.55 | 0.71* |
| | 89.69 to 99.24 | 36.60 to 55.65 | 1.50 to 2.13 | 0.03 to 0.25 | 47.80 to 64.78 | 84.88 to 98.86 | 0.64 to 0.77 |
| R5+S | 95.12* | 72.57* | 3.47 | 0.07 | 71.56 | 95.35 | 0.84* |
| | 87.98 to 98.66 | 63.37 to 80.54 | 2.56 to 4.70 | 0.03 to 0.18 | 62.12 to 79.79 | 88.52 to 98.72 | 0.78 to 0.89 |

+S means combining with the results of S-Detect.

*The integrated results of the residents and S-Detect were significantly different with the original ones, with p value <0.001.

†The integrated results of the residents and S-Detect were significantly different with the original ones, with p value <0.05.

AUC, area under the receiver operating characteristics curve; NLR, negative likelihood ratio; NPV, negative predictive value; PLR, positive likelihood ratio; PPV, positive predictive value; R, resident; SE, sensitivity; SP, specificity.

residents-in-training were integrated with those of the S-Detect in category 4a lesions. We compared the results of S-Detect and those of the residents for each lesion. If the lesion was diagnosed as category 4a by the residents but possibly benign by S-Detect, the results of S-Detect were adopted, thus downgrading category 4a lesions to the possibly benign group. Due to the high sensitivity of the residents presented in the preliminary experiments, we did not change the category 3 lesions when they were classified as possibly malignant by S-Detect. In addition, the rest of the classifications made by the residents remained unchanged.

Diagnostic performances of the integrated results were calculated and compared with the original results of the residents without S-Detect. Inter-rater variability before and after integration with S-Detect was assessed using intraclass correlation coefficient (ICC).

Statistical analysis

The diagnostic performances of the residents, S-Detect and the integrated results of the residents and S-Detect for category 4a lesions were evaluated using sensitivity, specificity, positive likelihood ratio, negative likelihood ratio, positive predictive value, negative predictive value, receiver operating characteristics (ROC) curve and area under the receiver operating characteristics curve (AUC). In addition, 2×2 contingency tables were delineated to measure these indicators. We performed comparisons of sensitivity and specificity between residents using the χ^2 test. The AUC values were compared using the Z test.

ICC with 95% CI was calculated to evaluate the inter-rater variability of multiple raters. In this study, each subject was rated by the raters, and ICC was deemed the absolute agreement of the raters, as the systematic differences among the raters were relevant. ICC value was interpreted as follows:

- ▶ Poor agreement: $ICC < 0$.
- ▶ Slight agreement: $0 < ICC < 0.20$.
- ▶ Fair agreement: $0.20 < ICC < 0.40$.
- ▶ Moderate agreement: $0.40 < ICC < 0.60$.
- ▶ Substantial agreement: $0.60 < ICC < 0.80$.
- ▶ Perfect agreement: $0.80 < ICC < 1$.

Statistical significance was considered when the p value was less than 0.05. SPSS V.21.0 software and MedCalc V.15 (MedCalc Software, Ghent, Belgium) were used in the study.

RESULTS

A total of 195 focal breast lesions, including 82 malignant lesions and 113 benign lesions, from 195 consecutive patients (mean age, 45.7 years; median age, 45.0 (15–82) years) who were referred to the medical centre were consecutively enrolled. The detailed pathological results and BI-RADS classifications are presented in [table 1](#).

The diagnostic performances of S-Detect and the five residents, and the comparisons of sensitivity, specificity and AUC between S-Detect and the residents, are listed

in [table 2](#). [Table 2](#) highlights that the residents had high sensitivity but evidently low specificity in classifying breast lesions. All residents showed a relatively high sensitivity (92.68%–100.00%). The specificity of S-Detect (77.88%) was higher than that of R2–R5 (19.47%–48.67%), with p value < 0.05 . The AUC value of S-Detect (0.82) was significantly higher than those of the five residents (0.62–0.74), with p value < 0.05 for all residents, as shown in [table 2](#). In this study, S-Detect had overall better diagnostic performance than the residents-in-training with limited breast US experiences.

The number of downgraded lesions that were classified as category 4a lesions by the residents but possibly benign by S-Detect is listed in [table 3](#). A total of 24, 31, 38, 32 and 42 identified as BI-RADS 4a lesions by R1, R2, R3, R4 and R5 were downgraded as possibly benign lesions by S-Detect, respectively. Among these downgraded lesions, 24, 28, 35, 30 and 40 lesions were proved to be pathologically benign, respectively, and 0, 3, 3, 2 and 2 downgraded lesions were malignant, respectively. A typical case of S-Detect-downgraded 4a lesion is shown in [figure 1](#). The mass was found in a 41-year-old woman and verified as a benign phyllodes tumour on histopathology. It was classified as BI-RADS 4a by four of the residents, based on its slightly irregular shape on greyscale US and abundant intratumorous vessels on colour Doppler US. S-Detect provided a diagnosis of possibly benign and downgraded it accurately.

The sensitivity of the integrated results remained at a relatively high level (92.68%–100.00%). The specificities of all residents significantly improved after using the results of S-Detect (46.02%–76.11%), with a p value < 0.001 for all residents. The ROC curves of the five residents, S-Detect and the residents combined with S-Detect are presented in [figure 2](#). From the ROC curves of the residents, we could determine that the curve was elevated at the top left after combination with S-Detect. Additionally, the AUC value of the residents with S-Detect had an evident increase (0.71–0.85), with statistical significance ($p < 0.001$), indicating improvement in the overall diagnostic performance of the five residents ([table 2](#)).

To evaluate the interobserver variability among the five residents, we calculated the ICC value of the integrated results and the original results. Systematic differences among the five raters were found to be relevant after analysis of variance ($p < 0.05$), and the ICC was regarded as a measure of absolute agreement. The single measure of ICC of the five residents increased from 0.480 (0.415–0.549) to 0.643 (0.586–0.700) after integration with the results of S-Detect, indicating that the agreement level increased from moderate to substantial.

DISCUSSION

US is one of the most commonly used modalities in breast imaging. As a convenient and cost-effective imaging method, US has played an essential role in the detection and evaluation of breast lesions in many countries,

Table 3 Downgraded 4a lesions by S-Detect

| | Total number of 4a lesions | Downgraded lesions | | Histologically malignant | Histologically benign |
|----|----------------------------|--------------------|--------------------|--------------------------|-----------------------|
| R1 | 32 | 24 | S-Detect malignant | 3 | 5 |
| | | | S-Detect benign | 0 | 24 |
| R2 | 37 | 31 | S-Detect malignant | 0 | 6 |
| | | | S-Detect benign | 3 | 28 |
| R3 | 52 | 38 | S-Detect malignant | 5 | 9 |
| | | | S-Detect benign | 3 | 35 |
| R4 | 39 | 32 | S-Detect malignant | 2 | 5 |
| | | | S-Detect benign | 2 | 30 |
| R5 | 55 | 42 | S-Detect malignant | 3 | 10 |
| | | | S-Detect benign | 2 | 40 |

R, resident.

including in China.²⁸ However, despite the promotion of BI-RADS lexicon, operator dependence and interobserver variability remain the major limitations of US.^{8–11} The performance of the BI-RADS lexicon can be largely affected by the clinical experiences of the operators. The specificity of a resident-in-training has been reported to be significantly inferior to that of a high-level radiologist, when using the BI-RADS lexicon in the assessment of breast lesions.¹⁰ As a result, methods to enhance the diagnostic efficiency of inexperienced readers and to decrease the interobserver variability in breast US findings are in demand.

CAD systems have emerged as a powerful tool for medical imaging with the dramatic advancement of artificial intelligence technology.¹⁷ The feasibility of using CAD systems to aid in the diagnosis of breast lesions has been verified by previous studies.^{29,30} S-Detect, a dedicated CAD software, was constructed on a deep learning algorithm and trained by large clinical databases and was integrated into a high-end US unit. The diagnostic process of S-Detect is free from the interference of man-identified features. The potential use of S-Detect to assist doctors in improving diagnostic performance, especially those who lack experience, has been elucidated in previous studies. Choi *et al* and Cho *et al*^{31,32} verified that the diagnostic performance of inexperienced readers could be improved with the help of S-Detect. Di Segni *et al*²² also suggested that S-Detect could serve as a teaching tool for residents-in-training to improve accuracy in diagnosing breast lesions.

According to the results of our study, S-Detect was distinguished by its high specificity, compared with that of the five residents-in-training with limited US experience, who presented remarkable sensitivity but low specificity.

Therefore, we speculated that S-Detect could help in improving residents' specificity. Breast lesions classified into BI-RADS 4a were defined as having low suspicion for malignancy. In the clinical setting, category 4a is a relatively complicated subgroup of the BI-RADS classifications, of which the malignant rate is 3%–10% and the positive predictive value is 6%.¹⁶ In this study, the ratio of malignancy in 4a lesions classified by the five residents was 9.38%, 8.11%, 15.38%, 10.25% and 9.09%, respectively, most of which were within the range defined by the guidelines. Most category 4a lesions are benign, but may undergo unnecessary biopsies. To better address the overtreatment of 4a lesions, new modalities, such as elastography, have been put into clinical use to lower the false-positive rate.^{33,34} In our study, 24 out of 32, 31 out of 37, 38 out of 52, 32 out of 39, and 42 out of 55 BI-RADS 4a lesions were downgraded by S-Detect, and most of the downgraded lesions proved to be benign (24 of 24, 28 of 31, 35 of 38, 30 of 32, and 40 of 42). Statistically significant improvement in the specificity and AUC was obtained for the residents after using S-Detect for category 4a lesions, suggesting that the dedicated CAD system might also provide additional diagnostic information. The CAD system could also be an effective method to downgrade benign category 4a lesions and reduce unnecessary biopsies. To note, the rate of malignancy of the CAD-downgraded 4a lesions in the study was higher than we expected, which was 0%, 9.68%, 7.89%, 6.25% and 4.76%, respectively. This implied that further improvement of S-Detect is necessary for clinical applications. Additionally, in this study, due to high sensitivity among the readers (96.34%–100%), S-Detect did not increase sensitivity in BI-RADS 3 lesions. More participants with different experiences in breast US are needed to further explore the value of S-Detect in enhancing sensitivity.

The ICC of the five residents improved after integration with S-Detect from a moderate level of agreement to a good level. This result verified that S-Detect could also be effective in decreasing interobserver variability in breast US among inexperienced raters. In the clinical practice, residents are required to undergo systematic training programmes before entering clinical work. S-Detect can act as a powerful adjunctive tool to audit the diagnoses made by inexperienced US readers. Notably, the workflow of S-Detect is less time-consuming than that of the double reading process. In addition, the US features extracted by S-Detect are displayed for readers, providing a useful reference for residents to learn the images case by case; thus, S-Detect may possess potential value in the training of inexperienced US readers.

There were several limitations to this study. First, the underestimation of the performance of the residents should be mentioned. In the regular US examination flow, radiologists often evaluate a breast lesion according to overall diagnostic information. Apart from dynamic real-time US images, medical history and mammography results are taken into consideration, while in this study only static images were provided for residents to classify. Second, S-Detect makes classification on the basis of one slice of the lesion, and an opposite diagnosis may occur when selecting different sections of the lesion. In this study, we used the maximal cross section of one lesion for automated diagnosis to make the methods more repeatable. Also, the good performance of S-Detect was guaranteed by high-quality US images used for classification, which were collected by the experienced radiologist who participated in the study. In real clinical settings, the diagnostic performance of S-Detect may decrease when substandard images were used for analysis in different medical centres. Therefore, high-quality US slices of a lesion acquired by standard procedure are necessary for future studies and clinical use of S-Detect. Lastly, this study was conducted in a single centre with only five residents participating. More radiologists at different levels of breast US and from other medical centres should be involved to establish the role of S-Detect in clinical application.

CONCLUSION

In this study, S-Detect had better diagnostic performance in classifying breast lesions than the five residents. After category 4a lesions were reclassified by S-Detect, the diagnostic performances of the residents significantly improved, with higher specificity but without sacrificing the sensitivity significantly. It is promising for S-Detect to improve the specificity of inexperienced readers and avoid unnecessary biopsies of category 4a lesions.

Contributors CZ arranged the data, performed the statistical analysis and wrote the manuscript. MX, HL, MW, HW, JZ and CZ collected the clinical and imaging data and performed the imaging assessments. YJ and QZ designed the study and participated in editing the manuscript.

Funding This work was funded by the CAMS Innovation Fund for Medical Sciences (2017-12M-1-006), the 2016 Peking Union Medical College education and teaching

reform project (2016zlgc0113), and the Fundamental Research Funds for the Central Universities (2017320002).

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Ethics approval for the study was obtained from the Institutional Review Board of Peking Union Medical College Hospital.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement The data sets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Chenyang Zhao <http://orcid.org/0000-0002-0618-2381>

REFERENCES

- 1 Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. *CA Cancer J Clin* 2015;65:5–29.
- 2 Ahmad A. Breast cancer statistics: recent trends. *Adv Exp Med Biol* 2019;1152:1–7.
- 3 Dubey AK, Gupta U, Jain S. Breast cancer statistics and prediction methodology: a systematic review and analysis. *Asian Pac J Cancer Prev* 2015;16:4237–45.
- 4 Tabár L, Vitak B, Chen HH, *et al*. The Swedish two-county trial twenty years later. updated mortality results and new insights from long-term follow-up. *Radiol Clin North Am* 2000;38:625–51.
- 5 Nelson HD, Fu R, Cantor A, *et al*. Effectiveness of breast cancer screening: systematic review and meta-analysis to update the 2009 U.S. preventive services Task force recommendation. *Ann Intern Med* 2016;164:244–55.
- 6 Brem RF, Lenihan MJ, Lieberman J, *et al*. Screening breast ultrasound: past, present, and future. *AJR Am J Roentgenol* 2015;204:234–40.
- 7 Shen S, Zhou Y, Xu Y, *et al*. A multi-centre randomised trial comparing ultrasound vs mammography for screening breast cancer in high-risk Chinese women. *Br J Cancer* 2015;112:998–1004.
- 8 Abdullah N, Mesurolle B, El-Khoury M, *et al*. Breast imaging reporting and data system lexicon for US: interobserver agreement for assessment of breast masses. *Radiology* 2009;252:665–72.
- 9 Elverici E, Zengin B, Nurdan Barca A, *et al*. Interobserver and intraobserver agreement of sonographic BIRADS lexicon in the assessment of breast masses. *Iran J Radiol* 2013;10:122–7.
- 10 Lee YJ, Choi SY, Kim KS, *et al*. Variability in observer performance between faculty members and residents using breast imaging reporting and data system (BI-RADS)-ultrasound, fifth edition (2013). *Iran J Radiol* 2016;13:e28281.
- 11 Park CS, Kim SH, Jung NY, *et al*. Interobserver variability of ultrasound elastography and the ultrasound BI-RADS lexicon of breast lesions. *Breast Cancer* 2015;22:153–60.
- 12 Rao AA, Feneis J, Lalonde C, Ojeda-Fournier H: A Pictorial Review of Changes in the BI-RADS. In: *Radiographics: a review publication of the Radiological Society of North America, Inc.* 5 edn, 2016: 623–39.
- 13 Radiology ACO. *The American College of radiology breast imaging reporting and data system (BI-RADS)*, 2003.
- 14 Baert AL. *Breast imaging reporting and data system (BI-RADS)*, 2013.
- 15 Youk JH, Jung I, Yoon JH, *et al*. Comparison of inter-observer variability and diagnostic performance of the fifth edition of BI-RADS for breast ultrasound of static versus video images. *Ultrasound Med Biol* 2016;42:2083–8.
- 16 Lazarus E, Mainiero MB, Schepps B, *et al*. BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value. *Radiology* 2006;239:385–91.
- 17 Dromain C, Boyer B, Ferré R, *et al*. Computed-aided diagnosis (CAD) in the detection of breast cancer. *Eur J Radiol* 2013;82:417–23.
- 18 Chang R-F, Wu W-J, Moon WK, *et al*. Improvement in breast tumor discrimination by support vector machines and speckle-emphasis texture analysis. *Ultrasound Med Biol* 2003;29:679–86.
- 19 Chen C-M, Chou Y-H, Han K-C, *et al*. Breast lesions on sonograms: computer-aided diagnosis with nearly setting-independent features and artificial neural networks. *Radiology* 2003;226:504–14.



- 20 Yap MH, Pons G, Marti J, *et al.* Automated breast ultrasound lesions detection using Convolutional neural networks. *IEEE J Biomed Health Inform* 2018;22:1218–26.
- 21 Becker AS, Mueller M, Stoffel E, *et al.* Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. *Br J Radiol* 2018;91:20170576.
- 22 Di Segni M, de Soccio V, Cantisani V, *et al.* Automated classification of focal breast lesions according to S-detect: validation and role as a clinical and teaching tool. *J Ultrasound* 2018;21:105–18.
- 23 Kim K, Song MK, Kim E-K, *et al.* Clinical application of S-Detect to breast masses on ultrasonography: a study evaluating the diagnostic performance and agreement with a dedicated breast radiologist. *Ultrasonography* 2017;36:3–9.
- 24 Choi J-H, Kang BJ, Baek JE, *et al.* Application of computer-aided diagnosis in breast ultrasound interpretation: improvements in diagnostic performance according to reader experience. *Ultrasonography* 2018;37:217–225.
- 25 Bartolotta TV, Orlando A, Cantisani V, *et al.* Focal breast lesion characterization according to the BI-RADS US lexicon: role of a computer-aided decision-making support. *Radiol Med* 2018;123:498–506.
- 26 Itoh A, Ueno E, Tohno E, *et al.* Breast disease: clinical application of US elastography for diagnosis. *Radiology* 2006;239:341–50.
- 27 Kim HJ, Kim SM, Kim B, *et al.* Comparison of strain and shear wave elastography for qualitative and quantitative assessment of breast masses in the same population. *Sci Rep* 2018;8:6197.
- 28 Guo R, Lu G, Qin B, *et al.* Ultrasound imaging technologies for breast cancer detection and management: a review. *Ultrasound Med Biol* 2018;44:37–70.
- 29 Drukker K, Giger ML, Metz CE. Robustness of computerized lesion detection and classification scheme across different breast US platforms. *Radiology* 2005;237:834–40.
- 30 Drukker K, Giger ML, Vyborny CJ, *et al.* Computerized detection and classification of cancer on breast ultrasound. *Acad Radiol* 2004;11:526–35.
- 31 Cho E, Kim E-K, Song MK, *et al.* Application of computer-aided diagnosis on breast ultrasonography: evaluation of diagnostic performances and agreement of radiologists according to different levels of experience. *J Ultrasound Med* 2018;37:209–16.
- 32 Choi J-H, Kang BJ, Baek JE, *et al.* Application of computer-aided diagnosis in breast ultrasound interpretation: improvements in diagnostic performance according to reader experience. *Ultrasonography* 2018;37:217–25.
- 33 Koh J, Kim E-K, Kim MJ, *et al.* Role of elastography for downgrading BI-RADS category 4A breast lesions according to risk factors. *Acta Radiol* 2019;60:278–85.
- 34 Au FW-F, Ghai S, Moshonov H, *et al.* Diagnostic performance of quantitative shear wave elastography in the evaluation of solid breast masses: determination of the most discriminatory parameter. *AJR Am J Roentgenol* 2014;203:W328–36.